## SALES ANALYSIS ON BIG BASKET PRODUCTS DATA SET

**DOMAIN: ECOMMERCE WEBSITES** 

# TOOLS USED: PYTHON FOR DATA CLEANING AND PRE-PROCESSING APACHE PIG FOR DATA ANALYSIS

**DATASOURCE: KAGGLE** 

https://www.kaggle.com/hetulmehta/bigbasket-products-dataset

**SCHEMA:** 

PRODUCT, CATEGORY, SUB-CATEGORY, BRAND, TYPE, RATING, SALE PRICE, MARKET PRICE, PROFIT, PROFIT%

**SIZE: 22.45 MB** 

### **DATA CLEANING IN PYTHON**

- UNECCESSARY COLUMNS LIKE URLS, CODES WERE REMOVED
- NULL VALUES AND DUPLICATES WERE REMOVED
- MISSING VALUES IN RATING COLUMNS WE REPLACED WITH ITS MEDIAN.

**DETAILED EXPLANATION IN JUPYTER NOTEBOOK** 



### **ANALYSIS PART USING PIG**

## 1) LOAD, DESCRIBE AND DUMP OPERATIONS



**DUMPED DATA** 

```
kevingkevin-VirtualBox: -/Desktop

(Altond Oil, Beauty & Hygiene, Skin Care, Arona Magic, 300.0, 400.0, Aronatherapy, 2.0, 100.0, 33.2)

(3-1n.1 Face Wash, Beauty & Hygiene, Skin Care, Organic Harvest, 182.75, 215.0, Face Care, 4.0, 32.25, 17.0)

(Radiant Pearl Cellular Lightening | Facil kit, Caleauty & Hygiene, Skin Care, Lotus Herbais, 171.5, 245.0, Face Care, 4.5, 73.5, 42.9)

(Radiant Pearl Cellular Lightening | Facil kit, Care, Beauty & Hygiene, Skin Care, Lotus Herbais, 171.5, 245.0, Face Care, 4.5, 73.5, 42.9)

(Rodiant Pearl Cellular Lightening | Facil kit, Care, Beauty & Hygiene, Skin Care, Op. 40.7 acc Care, 13.3, 0.0, 0.0)

(Colid Crene Black), Beauty & Hygiene, Skin Care, Natures Essence, 70.0, 70.0, Face Care, 13.3, 0.0, 0.0)

(Colid Crene Black), Beauty & Hygiene, Skin Care, Arona Magic, 148.75, 175.0, Face Care, 3.0, 175.5, 17.6)

(Sunscreen Lotin - Carrot, Beauty & Hygiene, Skin Care, Arona Magic, 148.75, 175.0, Face Care, 3.0, 0.0)

(Breancather Essentia) Oil, Beauty & Hygiene, Skin Care, The Face Shop, 100.0, 100.0, Face Care, 4.7, 2.0, 0.0)

(Breancather Essentia) Oil, Beauty & Hygiene, Skin Care, Soul Tlover, 300.0, 450.0, Aronatherapy, 4.8)

(Calarocal Peac Loff Face Nath, Beauty & Hygiene, Skin Care, 174.10, 175.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170.0, 170
```

### **QUERIES:**

## 1) MAXIMUM RATED PRODUCTS: THAT IS PRODUCTS WITH RATING 5

```
grunt> temp1 = FILTER project by rating==5.0;
2021-10-03 10:18:01,491 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2 time(s).
grunt> dump temp1;
```

FROM THE OUTPUT WE CAN SEE THAT THE HIGHEST RATED PRODUCTS ARE: DRINKS(TEA), SANITARY PRODUCTS AND PRODUCTS FOR BABIES LIKE TISSUES, DIAPERS ETC.

## 2) MINIMUM RATED PRODUCTS: THAT IS PRODUCTS WITH RATING 1

```
grunt> temp2 = FILTER project by rating==1.0;
2021-10-03 10:20:55,013 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPL
ICIT_CAST_TO_DOUBLE 3 time(s).
grunt> dump temp2
```

```
(Acmokilan-Rs Gel Face Wash, Beauty & Hygiene, Skin Care, Vegetal, 228.0, 228.0, Face Care, 1.0, 0.0, 0.0)
(Massage Oll - Aprilcol Kernel Oll, Beauty & Hygiene, Skin Care, Alone Ved, 472.5, 522.0, Aromatherapy, 1.0, 52.5, 11.1)
(Glow only Black Creen, Beauty & Hygiene, Skin Care, Clubus, 192.0, 248.0, 140.0, 23.0)
(Glow only Black) (Green, Beauty & Hygiene, Skin Care, Clubus, 192.0, 248.0, 140.0, 23.0)
(Getton) Advanced Anni Aging Face Mixt, Beauty & Hygiene, Skin Care, Stockance, 699.0, 699.0, 7ace Care, 1.0, 0.0, 0.0)
(Coccount Oll, Beauty & Hygiene, Skin Care, Stockance, 699.0, 699.0, 7ace Care, 1.0, 0.0, 0.0)
(Coccount Oll, Beauty & Hygiene, Skin Care, Stockance, 699.0, 699.0, 7ace Care, 1.0, 0.0, 0.0)
(Goody Indiana, 10.100, 180.0)
(Body Massage Oll - Zasnine & Nargis, Beauty & Hygiene, Skin Care, Orleant & Stotiance, 699.0, 899.8, Aromatherapy, 1.0, 0.0)
(Body Massage Oll - Zasnine & Nargis, Beauty & Hygiene, Skin Care, Orleant & Stotiance, 699.0, 899.8, Aromatherapy, 1.0, 0.0)
(Goody Massage Oll - Zasnine & Nargis, Beauty & Hygiene, Skin Care, Orleant & Stotiance, 699.0, 899.8, Aromatherapy, 1.0, 0.0)
(Goody Massage Oll - Zasnine & Nargis, Beauty & Hygiene, Skin Care, Organic Harver, 1311.6, 410.0, Anomatherapy, 1.0, 0.0, 0.0)
(Gracial Kit. - Pearl, Beauty & Hygiene, Skin Care, Aroma Nagis, C. 900.0, 1150.0, 7ace Care, 1.0, 400.0, 400.0)
(Faccial Kit. - Pearl, Beauty & Hygiene, Skin Care, Aroma Nagis, C. 900.0, 1150.0, 7ace Care, 1.0, 400.0, 400.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Andreant, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Mixter, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Andreant, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Mixter, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Mixter, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Mixter, 1910.0)
(Faccial Roby Mixt - Frankincene, Beauty & Hygiene, Skin Care, Mixter, 1910.0)
(Faccial
```

FROM THE OUTPUT WE CAN SEE THAT THE LEAST RATED PRODUCTS ARE: BEAUTY, SKIN-CARE AND HYGEIENE TYPE PRODUCTS AND SOME FOOD ITEMS.

### 3) NUMBER OF PRODUCTS PER BRAND:

```
grunt> gbrand = GROUP project by brand;
2021-10-03 10:26:01,147 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 3 time(s).
grunt> describe gbrand;
gbrand: {group: chararray,project: {(product: chararray,category: chararray,sub_category: chararray,brand: chararray,sale_price: float,market_price: i
itpercentage: float)}}
grunt> cnt_prod = foreach gbrand generate group , COUNT(project);
2021-10-03 10:27:11,870 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 3 time(s).
grunt> dump cnt_prod;
```

```
(Z,2)
(DP,11)
(Fa,3)
(HE,1)
(Re,1)
(V2,2)
(iD,6)
(365,1)
(3B0,4)
(BSY,1)
(Cif,4)
(Cpp,1)
(DND,1)
(Dip,6)
(Ela,15)
(Eva,5)
(Exo,5)
(Fem, 17)
(GPC,4)
(Grb,5)
(HIT,6)
(IN2,8)
(Joy, 39)
(KCL,3)
(Kvg,1)
(Lux,9)
(MOM, 10)
(MTR,65)
(MVS,2)
(Mud, 1)
(NSH,1)
(Neo,17)
```

FROM THE OUTPUT WE CAN SEE BRAND WISE COUNT OF PRODUCTS.

BRANDS like FRESHO, BBHOME, COLGATE ETC. HAVE MORE NO. OF PRODUCTS.

## 4) TOTAL PROFIT MADE BY EACH BRAND

```
grunt> gbrand = GROUP project by brand;
2021-10-03 10:44:34,012 [main] WARN org.apache
grunt> mpb = foreach gbrand {
>> ordp = order project by Profit desc;
>> maxpro = limit ordp 1;
>> generate group , SUM(project.Profit);
>> };
2021-10-03 10:45:14,928 [main] WARN org.apache
grunt> dump mpb;
```

```
(Al' Fez,0.0)
(Alkalen,0.0)
(All Out,94.0)
(Alvizia,476.0400085449219)
(Ambipur,0.0)
(Amrutha,7.400000095367432)
(Aramusk,445.0)
(Azafran,1528.9499998092651)
(BB Home,841.1000018119812)
(Bacto-V,0.0)
(Bagrrys,297.25)
(Bambino,0.0)
(Batiste,0.0)
(Battler,0.0)
(Beanies,140.0)
(Befikar,0.0)
(Bentley,2970.0)
(Bio-Oil,229.0)
(Bionova,0.0)
(Bolst's,0.0)
(Brimune,0.0)
(Britte ,0.0)
(Brother,38.75)
(Buffet ,0.0)
(Bvlgari,0.0)
(Cadbury,249.31999871134758)
(CamPure,0.0)
(Canopus,0.0)
(Carmesi,0.0)
(Cartini,57.99999952316284)
(Chaayos,0.0)
(Charmis,64.9000015258789)
(Cheetos,0.0)
(Cinthol,326.9600028991699)
(Clenz U,0.0)
(Closeup,90.0)
(Cocojal,0.0)
(Colgate,2918.3999996185303)
(Comfort,35.100000858306885)
(D-Alive,3.5)
(Daadis ,12.699999809265137)
(Daarzel,91.35000038146973)
(Delishh,4.799999952316284)
(Dexolac,0.0)
(Dheepam,0.0)
(Diamond,1.0)
(Dimples,0.0)
(Doritos,17.0)
```

TOTAL PROFIT MADE BY EACH BRAND IN RUPEES IS SHOWN IN THE OUTPUT.

## 5) AVERAGE RATING PER BRAND

```
grunt> gbrand = GROUP project by brand;
2021-10-03 11:16:11,170 [main] WARN org.apache.pig.newplan.BaseOperatorPlan -
grunt> e = foreach gbrand generate group , ROUND(AVG(project.rating));
2021-10-03 11:16:24,271 [main] WARN org.apache.pig.newplan.BaseOperatorPlan -
grunt>
```

### **OUTPUT:**

```
(Keya,4)
(Kiwi,4)
(LISO,1)
(Lays,4)
(MOOZ,4)
(Mars,5)
(Maxi,4)
(Maxo,4)
(Maya,)
(Moov,4)
(NOCD,4)
(Nice,5)
(Nike,4)
(Nova,3)
(Nyle,4)
(OGMO,3)
(Olay,4)
(On1y,4)
(Ongs,4)
(Open,4)
(Opex,4)
(Oreo,4)
(PIPO,4)
(Plum,4)
(Polo,)
(Post,4)
(Pril,4)
(Pure,4)
(QIPS,5)
(QRAA,4)
(RYCA,)
(Rage,)
(Real, 4)
(Red",)
(Rica,4)
(SOVI,5)
(Seni,4)
(Smac,4)
(Sofy,4)
(Solo,4)
(Tagz,4)
(Tang,4)
```

AVERAGE RATING PER BRAND IS SEEN IN THE OUTPUT I.E OUTPUT RATINGS IN THE RANGE(1-5)

## BRANDS LIKE OREO ,TANG HAVE AVERAGE RATING 4 AND BRANDS LIKE MARS AND NICE HAVE AVERAGE RATING 5

## 6) MOST PROFITABLE PRODUCT

```
grunt> h1 = group project all;
2021-10-03 12:04:47,111 [main] WARN org.apache.pig.newplan.Bagrunt> I = foreach h1 generate MAX(project.Profit) as max;
2021-10-03 12:05:41,872 [main] WARN org.apache.pig.newplan.Bagrunt> x = filter project by Profit==I.max;
2021-10-03 12:06:09,087 [main] WARN org.apache.pig.newplan.Bagrunt> dump x;
```

### **OUTPUT:**

```
2021-10-03 12:06:15,550 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUti
("Premium Cloth Dryer/Drying Stand - Foldable, Silver, BB 1434","Kitchen,,,DP,4649.0,8969.0,)
grunt>
```

FROM THE OUTPUT WE CAN SEE THE MOST PROFITABLE
PRODUCT IS PREMIUM CLOTH DRYER FROM THE BRAND DP

## 7) LEAST PROFITABLE PRODUCTS

```
grunt> h1 = group project all;

Thunderbird Mail 2:00,948 [main] WARN org.apache.pig.newplan.g

Thunderbird Mail 2:00,948 [main] WARN org.apache.pig.newplan.g

2021-10-03 12:24:49,976 [main] WARN org.apache.pig.newplan.g

grunt> x = filter project by Profit==I.min;

2021-10-03 12:27:27,121 [main] WARN org.apache.pig.newplan.g

grunt> dump x;
```

```
Citches Districtant & serveshing Spay, Clearing & Nussehind, of Europea Claimers, airlou, 175.8, 2015 interior Spring Clearer, Act, 6.0, 6.0)

(Money Floor (Learer, See, 300 ml. a) blue men floor (Learer, Clearer, Act, 6.0, 6.0)

(Black Phenyl, Clearing & Household, all Purpose Clearers, permit Chek, 48.0, 48.0, 48.0, 4100 ms of the Clearers, 4.1, 6.0, 6.0)

(Permiter (Learer) (
```

FROM THE OUTPUT WE CAN SEE THE PRODUCTS WITH 0
RUPEES PROFIT AND THESE ARE MOSTLY THE PRODUCTS
THAT FALL IN THE CATEGORY OF HOUSEHOLD/CLEANING ETC.

## 8) MOST EXPENSIVE PRODUCT

```
grunt> h1 = group project all;
2021-10-03 12:39:23,643 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning
IMPLICIT_CAST_TO_DOUBLE 4 time(s).
grunt> I = foreach h1 generate MAX(project.sale_price) as exp;
2021-10-03 12:39:50,249 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning
IMPLICIT_CAST_TO_DOUBLE 4 time(s).
grunt> x = filter project by sale_price==I.exp;
2021-10-03 12:40:09,023 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning
IMPLICIT_CAST_TO_DOUBLE 4 time(s).
grunt> dump x;
```

### **OUTPUT:**

```
- Total input paths to process : 1
(Epilator SE9-9961 Legs-Body-Face,Beauty & Hygiene,Feminine Hygiene,Braun,8184.44,10769.0,Hair Rem, oval,4.1,2584.56,31.6)
grunt>
```

FROM THE OUTPUT WE CAN SEE EPILATOR IS THE MOST EXPENSIVE PRODUCT AS PER SALE PRICE.

## 9) CHEAPEST PRODUCT

```
kevin@kevin-VirtualBox: ~/Desktop Q = - □  

2021-10-03 12:48:00,034 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warnin g IMPLICIT_CAST_TO_DOUBLE 4 time(s).

grunt> I = foreach h1 generate MIN(project.sale_price) as cheap;

2021-10-03 12:48:17,939 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warnin g IMPLICIT_CAST_TO_DOUBLE 4 time(s).

grunt> x = filter project by sale_price==I.cheap;

2021-10-03 12:48:25,078 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warnin g IMPLICIT_CAST_TO_DOUBLE 4 time(s).

grunt> dump x;
```

### **OUTPUT:**

```
Curry Leaves, Fruits & Vegetables, Herbs & Seasonings, Fresho, 2.45, 15.0, Indian & Exotic Herbs, 4.1, 12.55, 512.2)
grunt>
```

## FROM THE OUTPUT THE CHEAPEST PRODUCT IS FOUND TO BE CURRY LEAVES

## 10) NUMBER OF PRODUCTS PER TYPE

```
kevin@kevin-VirtualBox: ~/Desktop Q = _ _ _ \textbf{Sequence} \textbf{Q} \textbf{Seq} = _ _ _ \textbf{Seq} \textbf{Seq} \textbf{Seq} = _ _ _ _ _ \textbf{Seq} \textbf{Seq} = _ _ _ _ \textbf{Seq} \textbf{Seq} \textbf{Seq} = _ _ _ _ \textbf{Seq} \textbf{Seq} \textbf{Seq} = _ _ _ \textbf{Seq} \textbf{Seq} \te
```

```
(Antiseptics & Bandages,25)
(Baby Dishes & Utensils,7)
(Cooking & Baking Needs,5)
(Cream & Cheese Spreads,14)
(Electronic Accessories,12)
(Exam Pads & Pencil Box,21)
(Exotic & Flavoured Tea,47)
(Extra Virgin Olive Oil,50)
(Flavoured & Other Oils,51)
(Floor & Other Cleaners,113)
(Floss & Tongue Cleaner,3)
(Fresh Salads & Sprouts,10)
(Gourmet Tea & Tea Bags,295)
(Hair & Scalp Treatment,67)
(Hand Wash & Sanitizers, 206)
(Heat & Eat Ready Meals,134)
(Moustache & Beard Care,69)
(Other Baby Accessories,3)
(Pet Food & Accessories,3)
(Shower Gel & Body Wash,125)
(Sports & Energy Drinks,19)
(Supplements & Proteins,207)
(Breakfast & Snack Mixes,201)
(Cream Biscuits & Wafers,94)
(Dishwash Bars & Powders,29)
(Glucose & Milk Biscuits,18)
(Gourmet Juices & Drinks,92)
(Maternity Personal Care,15)
(Regular & White Vinegar,5)
(Soap Cases & Dispensers,10)
(Appliances & Electricals,13)
(Balsamic & Cider Vinegar,35)
(Body Scrubs & Exfoliants,10)
(Disposable Cups & Plates,16)
(Flavoured & Greek Yogurt,22)
(Imported Oats & Porridge,51)
(International Chocolates,103)
(Namkeen & Savoury Snacks,265)
(Organic & Cold Press Oil,16)
(Tampons & Menstrual Cups,26)
(Dhatu Organics & Naturals,3)
(Dishwash Liquids & Pastes,39)
(Dry Shampoo & Conditioner,8)
(Face Masks & Safety Gears,34)
(Maternity Health Supplements,5)
(Disinfectant Spray & Cleaners,62)
```

FROM THE OUTPUT WE CAN SEE COUNT BY TYPE
FACE CARE, SHAMPOO, HANDWASH HAVE HIGH NUMBERS.

**CONCLUSION:** 

### **MOST PROFITABLE CATEGORIES ARE:**

### **BEAUTY & HYGIENE**

Comparatively fruits and vegetables and cleaning products are least profitable

So marketing and advertising or discounts can be added

And low rated products like cleaning and household should be also taken care using similar strategies in order to increase sales and profit.